文章编号:1001 - 9081(2009)03 - 0836 - 03

一种基于模糊聚类的构造进化树方法

李刚成[1],刘赞波[2],曾庆光[2]
(1. 湖南信息职业技术学院 信息工程系,长沙 410200; 2. 湖南大学 计算机与通信学院,长沙 410082)
(jt_lgc@126.com)

摘 要:各种生物之间的进化史可以通过构建进化树来讨论,因此进化树的研究成了一个研究热点。提出将利用 DNA 序列的 4D 表示所得相似矩阵视为模糊矩阵,再利用最大树法来构建进化树的方法。该方法不需要多序列比对,计算简单,实验验证了该方法的有效性。

关键词:DNA 序列;进化树; 模糊理论; 聚类分析; 最大树法
中图分类号: TP301.6;TP311.13 **文献标志码**:A

Method of constructing phylogenetic tree based on fuzzy clustering

LI Gang-cheng[1], LIU Zan-bo[2], ZENG Qing-guang[2]
(1. *Department of Information Engineering, Hunan College of Information, Changsha Hunan* 410200, *China*;
2. *School of Computer and Communication, Hunan University, Changsha Hunan* 410082, *China*)

Abstract: The evolutionary history of various species can be disscussed by constructing the phylogeny trees. Therefore, many scientists focus on the research of the phylogeny trees. In this paper, the authors used the similarity matrix computed by 4D representation of DNA sequence and regarded it as a fuzzy matrix, then used the maximal tree method to construct the phylogeny tree. This method need not sequence alignment and the computation is simple. The experiments demonstrate its validity.

Key words: DNA sequence; phylogenetic tree; fuzzy theory; cluster analysis; maximum tree method

0 引言

系统发生(phylogeny)是指一群有机体发生或进化的历史。系统发生树(phylogenetic tree),又称进化树(evolutionary tree),就是描述这一群有机体发生或进化顺序的拓扑结构,它可以用来研究不同物种间的进化关系,这一直是生物学的研究热点。

发生树推断就是一个根据某种标准,从给定的一组序列数据中推导出这些对象之间"最好"的系统发生树的过程。在生物领域内,待处理对象通常是生命机体、基因组和基因序列。用统计方法重建系统发生树分别独立地起始于形态学性状的数值分类法和分析基因频率数据的群体遗传学。在这些学科中发展起来的某些统计方法至今仍然用于分子数据的系统发生分析。现在最常用的推断系统发生树的方法可以分为两大类[1-2]:基于算法的方法和基于最优原则的方法。传统的用于进化树重构的距离测度包括 p 距离、Kimura 距离、Γ 距离等[1]。它们共同的特征是距离的计算都基于序列间的比对分析。比对分析对数据的要求严格,通常需要提取序列中的不同功能片段甚至要进行基因的预测[3]。同时,用于比对的计分矩阵也因为比对物种的不同和比对数据的不同而存在很大的经验性[1]。为了克服传统的基于序列比对的距离测度的不足,许多学者尝试用非比对的方法来比较 DNA 序列[4-6]。

目前最常见的基于算法的构建系统发生树的方法为距离法。在距离法[7]中,首先需要根据某种进化模型计算所有对象间的进化距离,然后根据不同的算法,从进化距离最短的开始依次聚类,利用距离方阵计算出最优树,或将分支长度之和最小化,获得最优树。

依据不同的聚类算法,距离法又有以下几种:使用算术平均的不加权的组对法[8](Unweighted Pair Group Method with Arithmetic mean, UPGMA)、Fitch-Margoliash 法[9]、邻接法(Neighbor Joining, NJ)[10],其中每种方法都有自己的优缺点。

基于最优原则的方法从数学角度讲就是在评价树的最优标准的基础之上,找到使得目标函数最优的树。目前最常用两种基于最优原则的方法有最大简约标准(Maximum Parsimony, MP)[11-12]和最大似然(Maximum Likelihood, ML)标准[13-14]。但是在分子数据大量积累的今天,往往所处理的数据量又是非常大的。并且,现在已经证明了构建 n 个对象的最大简约树和最大似然树[15]都是 NP 难的。因此针对两种方法,目前出现了许多改进算法。

上面我们提到的系统发生树构建方法都是基于算法的方法和基于最优原则的方法。聚类分析内容非常丰富,有系统聚类法、有序样品聚类法、动态聚类法、模糊聚类法、图论聚类法、聚类预报法等。本论文将要介绍的是一种利用模糊数学理论来构造进化树的方法。由于类与类之间存在模糊性,所以人们将模糊理论引入分类,从而产生了模糊聚类分析。

模糊聚类分析的方法很多,其中用得较多的有传递闭包法、最大树法、编网法、布尔矩阵法和动态直接聚类法[16]。这些聚类法有一个共同点,就是聚类依据是由原始数据所构造的模糊相似矩阵。聚类正确与否,完全取决于模糊相似矩阵。尽管模糊相似矩阵在模糊聚类分析中起决定性作用,但遗憾的是,模糊相似矩阵的构造方法不唯一。在讨论具体的方法之前,先简单介绍一些模糊数学的理论知识[16]。

1 模糊数学理论简介

定义 1 如果对于任意 $i = 1, 2, \cdots, m; j = 1, 2, \cdots, n$,都有 $r_{ij} \in [0,1]$,则称矩阵 $\boldsymbol{R} = (r_{ij})_{m\times n}$ 为模糊矩阵。

定义 2 若模糊关系 $\tilde{R} \in \mathscr{F}(X \times X)$ 满足:

收稿日期:2008 - 09 - 16;**修回日期**:2008 - 11 - 08。 **基金项目**:湖南省自然科学基金资助项目(07JJ5080;06JJ4076)。

作者简介:李刚成(1974 -),男,湖南邵阳人,讲师,硕士,主要研究方向:信息处理; 刘赞波(1981 -),男,湖南娄底人,硕士研究生,主要研究方向:信息处理; 曾庆光(1956 -),男,湖南长沙人,教授,硕士,主要研究方向:信息处理、网络技术。

1)自反性:$\tilde{R}(x,x) = 1$;

2)对称性:$\tilde{R}(x,y) = \tilde{R}(y,x)$;

3)传递性:$\tilde{R} \circ \tilde{R} \subseteq \tilde{R}$;

则称 $\tilde{R}$ 为 X 上的一个模糊等价关系。

定义 3　设论域 $U = \{x_1, x_2, \cdots, x_n\}$,$\boldsymbol{R} \in \mu_{n\times n}$,$\boldsymbol{I}$ 为单位矩阵,若 $\boldsymbol{R}$ 满足:

1)自反性:$\boldsymbol{I} \leq \boldsymbol{R}$　$(\Leftrightarrow r_{ij} = 1)$;

2)对称性:$\boldsymbol{R}^{\mathrm{T}} = \boldsymbol{R}$　$(\Leftrightarrow r_{ij} = r_{ji})$;

3)传递性:$\boldsymbol{R} \circ \boldsymbol{R} \quad \leq \boldsymbol{R}(\Leftrightarrow \bigvee_{k=1}^{n} (r_{ik} \wedge r_{kj}) \leq r_{ij})$;

则称 $\boldsymbol{R}$ 为模糊等价矩阵。

定义 4　若模糊关系 $\tilde{R} \in \mathfrak{S}(U \times U)$ 满足:

1)自反性:$\tilde{R}(x,x) = 1$;

2)对称性:$\tilde{R}(x,y) = \tilde{R}(y,x)$;

则称 $\tilde{R}$ 为 U 上的模糊相似关系。

定义 5　设论域 $U = \{x_1, x_2, \cdots, x_n\}$,$\boldsymbol{R} \in \mu_{n\times n}$,$\boldsymbol{I}$ 为单位矩阵,若 $\boldsymbol{R}$ 满足:

1)自反性:$\boldsymbol{I} \leq \boldsymbol{R}$　$(\Leftrightarrow r_{ij} = 1)$;

2)对称性:$\boldsymbol{R}^{\mathrm{T}} = \boldsymbol{R}$　$(\Leftrightarrow r_{ij} = r_{ji})$;

则称 R 为模糊相似矩阵。

定理 1　设 $\boldsymbol{R} \in \mu_{m\times m}$ 是模糊相似矩阵,则 $\forall k \in \mathbf{N}$,$\boldsymbol{R}^k$ 也是模糊相似矩阵。[16]

定理 2　设 $\boldsymbol{R} \in \mu_{m\times m}$ 是模糊相似矩阵,则存在一个最小自然数 $k(k \leqslant n)$,使得传递闭包 $t(\boldsymbol{R}) = \boldsymbol{R}^k$,对于一切大于 k 的自然数 l,恒有 $\boldsymbol{R}^l = \boldsymbol{R}^k$。此时,$t(\boldsymbol{R})$ 为模糊相似矩阵。[16]

由模糊数学理论知道,可以通过求传递闭包的方法来进行动态聚类,最后构建成我们的系统发生树。2006 年廖波在参考文献[17 - 18]中,就是利用传递闭包的方法来进行聚类分析,进而得到理想的系统发生树。

求传递闭包的方法在被聚类的物种数目比较多时,要把所建立的模糊相似关系“改造”成模糊等价关系是相当麻烦的。本文为了解决上面的问题,提出一种比较简便的方法——最大树法。

最大树法的基本思想:画出以被分类元素为顶点,以相似矩阵 $\boldsymbol{R}$ 的元素 r_{ij} 为权重的一棵最大的树,取定 $\lambda \in [0,1]$,砍断权重低于 λ 的枝,得到一个不连通的图,各个连通的分支便构成了 λ 水平上的分类。

最大树法进行模糊聚类分析的具体步骤如下:先画出被分类的所有元素,直接以模糊相似矩阵 $\boldsymbol{R}$ 中按 r_{ij} 由大到小的顺序依次把这些元素用直线连起来,并标上 r_{ij} 的数值。如果某一步使图中出现了回路,就不画这一步,依次走下一步,直到所有元素连通为止。这样就得到了一棵所谓的最大树(最大树不是唯一的,但不影响分类的结果)。然后,取定 λ 值($0 \leqslant \lambda \leqslant 1$),把 $r_{ij} < \lambda$ 的连接去掉,互相连通的元素归为一类,即可将元素分类。

算法思想如下:

1)T 的初始状态。只有 n 个顶点而无边的森林 $T = (V, \varnothing)$。

2)按边长递减的顺序选择 E 中的 $n - 1$ 安全边(u,v) 并加入 T,生成 MLT;

注意:安全边指两个端点分别是森林 T 里两棵树中的顶点的边。加入安全边,可将森林中的两棵树连接成一棵更大的树。

因为每一次添加到 T 中的边均是当前权值最大的安全边,MLT 性质也能保证最终的 T 是一棵最大生成树。

本算法特点如下:当前形成的集合 T 除最后的结果外,始终是一个森林。

Kruskal 算法的抽象描述如下:

```
KruskalMLT( G ){                          //求连通图 G 的一棵 MLT
  T = (V,∅) ; //初始化, T 是只含 n 个顶点不包含边的森林
  依权值的递减序对 E ( G )中的边排序,并设结果在 E[0, e - 1]中
  for( i = 0; i < e; i ++) {                    // e 为图中边总数
    取 E[0..e - 1) 中的第 i 条边(u,v);
    if u 和 v 分别属于 T 中两棵不同的树 then
      T = T ∪ {(u,v)};     //(u,v) 是安全边,将其加入 T 中
    if T 已是一棵生成树 then
      return T;
  }                                                    //endfor
  return T;
}
```

表 1　基于矩阵特征值向量端点之间欧氏距离的 11 种不同物种的相似性矩阵

Species	human	goat	gallus	opossum	lemur	mouse	rabbit	rat	bovine	gorilla	chimpanzee
human	0.0000	0.0260	0.1710	0.2660	0.0250	0.0500	0.1730	0.0980	0.0220	0.0180	0.0290
goat	0.0260	0.0000	0.1270	0.3620	0.0510	0.1260	0.2610	0.0360	0.0540	0.0580	0.0690
gallus	0.1710	0.1270	0.0000	0.8590	0.3060	0.4010	0.6860	0.0430	0.2990	0.2950	0.3340
opossum	0.2660	0.3620	0.8590	0.0000	0.1450	0.0980	0.0110	0.6260	0.1460	0.1480	0.1230
lemur	0.0250	0.0510	0.3060	0.1450	0.0000	0.0250	0.0820	0.1710	0.0030	0.0050	0.0020
mouse	0.0500	0.1260	0.4010	0.0980	0.0250	0.0000	0.0450	0.2820	0.0200	0.0140	0.0130
rabbit	0.1730	0.2610	0.6860	0.0110	0.0820	0.0450	0.0000	0.4890	0.0830	0.0830	0.0640
rat	0.0980	0.0360	0.0430	0.6260	0.1710	0.2820	0.4890	0.0000	0.1740	0.1780	0.2010
bovine	0.0220	0.0540	0.2990	0.1460	0.0030	0.0200	0.0830	0.1740	0.0000	0.0020	0.0030
gorilla	0.0180	0.0580	0.2950	0.1480	0.0050	0.0140	0.0830	0.1780	0.0020	0.0000	0.0030
chimpanzee	0.0290	0.0690	0.3340	0.1230	0.0020	0.0130	0.0640	0.2010	0.0030	0.0030	0.0000

2　实验过程

依据上面的模糊数学理论,我们就可以讨论文献[19 - 20]中得到的相似性矩阵,如表 1。表 1 明显满足相似性矩阵,可以经过简单改造后满足定义 1 和定义 5,那么此矩阵可看成是模糊相似矩阵。

直接从模糊相似矩阵出发,按照上面给出的最大树构建的方法,可以得到最大树,如图 1 和 2 所示。

表2 基于9维向量夹角余弦值的11种动物的相似性矩阵

Species	human	goat	opossum	mouse	gallus	lemur	rabbit	rat	gorilla	chimpanzee	bovine
human	1.000000	0.997298	0.994367	0.998526	0.997911	0.995595	0.997060	0.996658	0.999933	0.999682	0.998134
goat		1.000000	0.988730	0.993282	0.998453	0.994193	0.999038	0.995698	0.997999	0.998720	0.999828
opossum			1.000000	0.996382	0.990317	0.995613	0.990525	0.997366	0.993444	0.992466	0.990975
mouse				1.000000	0.993172	0.997256	0.995156	0.996849	0.997917	0.997460	0.995098
gallus					1.000000	0.991055	0.995890	0.994284	0.998425	0.998375	0.998247
lemur						1.000000	0.997445	0.999107	0.995270	0.995660	0.995936
rabbit							1.000000	0.997447	0.997491	0.998370	0.999496
rat								1.000000	0.996450	0.996567	0.997082
gorilla									1.000000	0.999877	0.998632
chimpanzee										1.000000	0.999227
bovine											1.000000

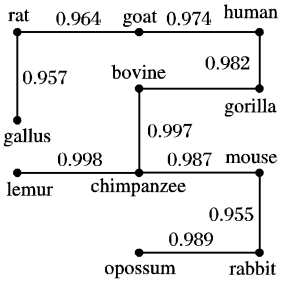

图1 基于表1求得的最大树

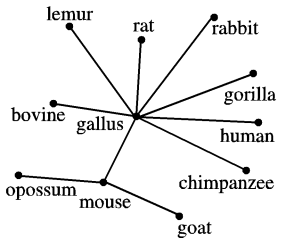

图2 基于表2求得的最大树

取定 $\lambda \in [0,1]$，砍断权重低于 λ 的枝，得到一个不连通的图，各个连通的分支便构成了 λ 水平上的分类。对于图1中，我们选取：$\lambda \in \{1.000, 0.998, 0.997, 0.989, 0.987, 0.982, 0.974, 0.964, 0.957, 0.955\}$。取 $\lambda = 1$，11个物种分为11类：{human}，{goat}，{gallus}，{opossum}，{mouse}，{rabbit}，{rat}，{bovine}，{gorilla}，{lemur}，{chimpanzee}；取 $\lambda = 0.998$，11个物种分为9类：{human}，{goat}，{gallus}，{opossum}，{mouse}，{rabbit}，{rat}，{bovine,gorilla}，{lemur,chimpanzee}；取 $\lambda = 0.997$，11个物种分为8类：{human}，{goat}，{gallus}，{opossum}，{mouse}，{rabbit}，{rat}，{bovine,gorilla,lemur,chimpanzee}；取 $\lambda = 0.989$，11个物种分为7类：{human}，{goat}，{gallus}，{rabbit, opossum}，{mouse}，{rat}，{bovine,gorilla,lemur,chimpanzee}；取 $\lambda = 0.987$，11个物种分为6类：{human}，{goat}，{gallus}，{rabbit, opossum}，{rat}，{mouse, bovine, gorilla, lemur, chimpanzee}；取 $\lambda = 0.982$，11个物种分为5类：{goat}，{gallus}，{rabbit, opossum}，{rat}，{mouse, bovine, gorilla, lemur, chimpanzee,human}；取 $\lambda = 0.974$，11个物种分为4类：{gallus}，{rabbit, opossum}，{rat}，{mouse, bovine, gorilla, lemur, chimpanzee,human,goat}；取 $\lambda = 0.964$，11个物种分为3类：{gallus}，{rabbit, opossum}，{mouse, bovine, gorilla, lemur, chimpanzee,human,goat,rat}；取 $\lambda = 0.957$，11个物种分为2类：{rabbit, opossum}，{mouse, bovine, gorilla, lemur, chimpanzee, human, goat, rat, gallus}；取 $\lambda = 0.955$，11个物种分为1类：{rabbit,opossum,mouse,bovine,gorilla,lemur,chimpanzee,human, goat,rat,gallus}。

根据最大树法得到的动态聚类图——11个物种基因序列的系统发生树，如图3。

同理，对于图2中，我们选取：$\lambda \in \{1.0000, 0.7214, 0.5838, 0.5831, 0.4844, 0.4681, 0.3797, 0.3785, 0.3635, 0.2749, 0.2746\}$，可以得到参考文献[20]中11个物种基因序列的系统发生树，如图4所示。

我们利用模糊聚类法构造的系统发生树与PHYLIP构造软件中NEIGHBOR程序构造的树大致是一致的[20]。由图3和4，我们可以观察到bovine，gorilla和chimpanzee三者很相似，而gallus和opossum跟其他物种的相似性较低，这与参考文献[19]中得到的结果也是一致的。当然利用模糊聚类的其他方法如传递闭包法、布尔矩阵法得到的进化树也是一样的[20]。

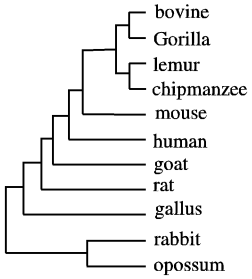

图3 基于图1的系统发生树

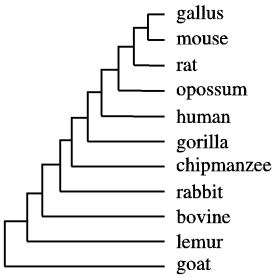

图4 基于图2的系统发生树

3 结语

本文在DNA序列4D表示的基础上，利用模糊理论构造进化树。与基于算法的方法和基于最优原则的方法相比，此方法的优点是不需要进行多序列比对和进化模型的建立，整个过程计算简单。我们利用最大树法构造的系统发生树与PHYLIP构造软件中NEIGHBOR程序构造的树是一致的。

参考文献：

[1] NEI M, KUMAR S. 分子进化与系统发生[M]. 吕宝忠, 钟扬, 高莉萍, 等译. 北京: 高等教育出版社, 2002.

[2] 郝柏林, 张淑誉. 生物信息学手册. 2版. 上海: 上海科学技术出版社, 2002: 200-204.

[3] MOUNT D W. Bioinformatics: sequence and genome analysis[M]. Cold Spring Harbor, NY: Cold Spring Harbor Laboratory Press, 2001: 337-342.

[4] VINGA S, ALMEIDAL J. Alignment-free sequence comparison: A review[J]. Bioinformatics, 2003, 19(4): 513-523.

[5] HAO BAILIN, QI JI, WANG BIN. Prokaryotic phylogeny based on complete genomes without sequence alignment[J]. Modern Physics Letters B, 2003, 17(2): 91-94.

[6] HAO BAILIN, QI JI. Prokaryote phylogeny without sequence alignment: from avoidance signature to composition distance[J]. Journal of Bioinformatics and Computational Biology, 2004, 2(1): 1-19.

[7] PAUPLIN Y. Direct calculation of a tree length using a distance matrix[J]. Journal of Molecular Evolution, 2000, 51(1): 41-47.

[8] THOMAS D. Example calculation of phylogenies: the UPGMA method [EB/OL]. [2008-09-01]. http://www.nmsr.org/upgma.htm.

[9] Chris. Fitch-Margoliash algorithm for calculating the branch lengths [EB/OL]. [2008-09-01]. http://www.bioinfo.rpi.edu/~bystrc/courses/biol4540/ecture12/sld002.htm.

（下转第842页）

图2中“CostOrder”代表文献[7-8]以代价为选择分裂属性衡量标准的方法,“InforCost”是本文考虑了代价和属性所含信息的方法,“Chi”方法[6]是一种使用卡方检验选择分裂属性的方法,试验中选用此方法进行比较的目的是因为卡方方法在文献[6]已经被证明优于信息熵的方法,但是文献[6]没有显示此方法是否在代价敏感决策树中也优于信息熵方法,因此,本文也构造相关试验来测试这个问题。

从实验结果可以看出:

1)当测试代价小时,三种方法误分类代价减少得很慢,但是本文方法对低测试代价没有其他两种方法敏感。

2)在相同的测试代价情况下,本文方法在四个数据中所产生的误分类错误代价都比另外两种方法低,即效果好。

3)在要降低相同的误分类代价情况下,本文方法所需的测试代价都比两种方法要少,而代价敏感分类的原则就是同等测试代价下误分类代价降低最多,这说明本文算法优于其他算法。

4)从四个试验结果可以看出,卡方算法均比信息熵算法效果要差。这说明虽然文献[6]已经证明在非代价敏感分类中,卡方方法优于信息熵方法,但是在代价敏感学习中却不是这样。

5)从图2可以发现,到了一定测试代价,误分类代价就减少得缓慢了,这个实验结果具有非常大的实用价值。比如,医疗过程中可以告诉患者,即使再多加钱,病情(即误分类代价)也不能得到成比例甚至不能再减轻,可以避免患者浪费钱,也可以避免患者受不良医务人员的欺骗;同时我们也发现,相对文献的方法,我们的方法花较少的代价就可以使误分类代价达到稳定值(因为这种分类问题通常是NP问题,没有最优解)。

6)从各图中所使用的测试代价可以看出,数据集Breast和数据集Heart有差不多个的属性,但是由于数据集事例不同,例如数据集Heart事例较少,所以实验的效果稍差,图中另外两个数据集Australia和数据集Voting也有同样的结果。

4 结语

本文利用每个属性所含的信息量和测试代价,提出了性价比的概念,并且综合了性价比和误分类减少这两个量,提出了一种新的分裂属性的选择方法。实验证明我们的方法优于文献[7-8]的方法和文献[6]的方法,并且从分析实验结果可以看出,本文方法有一定的实用价值。

参考文献:

[1] QUINLAN J R. Induction of decision trees[J]. Machine Learning, 1986, 1(1): 81-106.

[2] NUNEZ M. The use of background knowledge in decision tree induction[J]. Machine Learning, 1991, 6(3): 231-250.

[3] TAN M. Cost-sensitive learning of classification knowledge and its applications in robotics[J]. Machine Learning, 1993, 13(1): 7-33.

[4] LING C X, YANG Q. Decision trees with minimal costs[C]// Proceedings of 2004 International Conference on Machine Learning (ICML'2004). New York: ACM, 2004: 69.

[5] CHAI XIAOYONG, DENG LIN, YANG QIANG, *et al*. Test-cost sensitive naive Bayes classification [C]// Proceedings of The 2004 IEEE International Conference on Data Mining (ICDM'2004). Washington, DC: IEEE Computer Society, 2004: 51-58.

[6] 刘星毅. 一种新的决策树分类属性选择方法[J]. 计算机技术与发展, 2008, 18(5): 70-72.

[7] QIN ZHENGXIN, ZHANG SHICHAO, ZHANG CENGQI. Cost-sensitive decision trees with multiple cost scales [C]// Australian Conference on Artificial Intelligence, LNCS 3339. Berlin: Spring, 2005: 380-390.

[8] QIN ZHENXING, ZHANG CHENGQI, XIE XUEHUI, *et al*. Dynamic test-sensitive decision trees with multiple cost scales [C]// Proceedings of FSKD 2005, LNCS 3613. Berlin: Springer, 2005: 402-405.

[9] GASCHNIG J, DUDA R O, HART P E. Model design in the prospector consultant system for mineral exploration [C]// Expert Systems in the Microelectronic Age. Edinburgh, Scotland: Edinburgh University Press, 1979.

[10] BLADE C L, MERZ C J. UCI repository of machine learning databases[Z]. Irvine, CA: Department of Information and Computer Science, University of California, 1998.

[11] FAYYAD U M, IRANI K B. Multi-interval discretization of continuous-valued attributes for classification learning [C]// Proceedings of IJCAI. San Francisco: Morgan Kaufmann, 1993: 1022-1027.

[12] TURNEY P. Types of cost in inductive concept learning [C]// Proceedings of the Cost-Sensitive Learning Workshop at the 17th ICML-2000 Conference. San Francisco: Morgan Kaufmann, 2000: 15-21.

(上接第838页)

[10] SAITOU N, NEI M. The neighbor-joining method: A new method for reconstructing phylogenetic trees[J]. Molecular Biology and Evolution, 1987, 4(4): 406-425.

[11] SOBER E. Reconstructing the past: Parsimony, evolution and inference[M]. Cambridge: MIT Press, 1988.

[12] SOURDIS J, NEI M. Relative efficiencies of the maximum parsimony and distance matrix methods in obtaining the correct phylogenetic tree [J]. Molecular Biology and Evolution, 1988, 5(3): 298-311.

[13] HOLDER M, LEWIS P O. Phylogeny estimation: Traditional and Bayesian approaches[J]. Journal of Nature Reviews Genetics, 2003(4): 275-284.

[14] LI W-H. Evolutionary change of restriction cleavage sites and phylogenetic inference[J]. Genetics, 1986, 113(1): 187-213.

[15] ROCH S. A short proof that phylogenetic tree reconstruction by maximum likelihood is hard [J]. ACM Transactions on Computational Biology and Bioinformatics, 2006, 3(1): 92-94.

[16] 谢季坚, 刘承平. 模糊数学方法及其应用[M]. 3版. 武汉: 华中科技大学出版社, 2007.

[17] LIAO BO, SHAN XINZHOU, ZHU WEN, *et al*. Phylogenetic tree construction based on 2D graphical representation[J]. Chenical Physics Letters, 2006, 422(1/3): 282-288.

[18] WANG WEIPING, LIAO BO, WANG TIANMING, *et al*. A graphical method to construct a phylogenetic tree[J]. International Journal of Quantum Chemistry, 2006, 106(9): 1998-2005.

[19] 李刚成, 廖国华. 基于4D表示的DNA序列分析方法[J]. 科学技术与工程, 2008, 8(6): 1405-1409

[20] CAO ZHI, LIAO BO, LI RENFA. A group of 3 D graphical representation of DNA sequences based on dual nucleotides[J]. International Journal of Quantum Chemistry, 2008, 108(9): 1485-1490.